Biostatistics Series Module 2: Overview of Hypothesis Testing

This article has been cited by other articles in PMC.

Go to:

Hypothesis testing (or statistical inference) is one of the major applications of biostatistics. Much of medical research begins with a research question that can be framed as a hypothesis. Inferential statistics begins with a null hypothesis that reflects the conservative position of no change or no difference in comparison to baseline or between groups. Usually, the researcher has reason to believe that there is some effect or some difference which is the alternative hypothesis. The researcher therefore proceeds to study samples and measure outcomes in the hope of generating evidence strong enough for the statistician to be able to reject the null hypothesis. The concept of the P value is almost universally used in hypothesis testing. It denotes the probability of obtaining by chance a result at least as extreme as that observed, even when the null hypothesis is true and no real difference exists. Usually, if P is < 0.05 the null hypothesis is rejected and sample results are deemed statistically significant. With the increasing availability of computers and access to specialized statistical software, the drudgery involved in statistical calculations is now a thing of the past, once the learning curve of the software has been traversed. The life sciences researcher is therefore free to devote oneself to optimally designing the study, carefully selecting the hypothesis tests to be applied, and taking care in conducting the study well. Unfortunately, selecting the right test seems difficult initially. Thinking of the research hypothesis as addressing one of five generic research questions helps in selection of the right hypothesis test. In addition, it is important to be clear about the nature of the variables (e.g., numerical vs. categorical; parametric vs. nonparametric) and the number of groups or data sets being compared (e.g., two or more than two) at a time. The same research question may be explored by more than one type of hypothesis test. While this may be of utility in highlighting different aspects of the problem, merely reapplying different tests to the same issue in the hope of finding a P < 0.05 is a wrong use of statistics. Finally, it

is becoming the norm that an estimate of the size of any effect, expressed with its 95% confidence interval, is required for meaningful interpretation of results. A large study is likely to have a small (and therefore "statistically significant") *P* value, but a "real" estimate of the effect would be provided by the 95% confidence interval. If the intervals overlap between two interventions, then the difference between them is not so clear-cut even if *P* < 0.05. The two approaches are now considered complementary to one another.

Keywords: *Confidence interval*, *hypothesis testing*, *inferential statistics*, *null hypothesis*, *P value*, *research question*

Go to:
Much of statistics involves collecting, describing, and analyzing data that are subject to random variation. Descriptive statistics presents and summarizes collected data to characterize features of their distribution. Inferential statistics analyzes sample data in order to estimate or predict characteristics of the larger population from which the sample is drawn.

Biological phenomena are inherently variable and in this age of "evidence-based medicine" an understanding of such variation through statistical approaches is essential not only for the medical researcher who intends to draw inferences from his sample, but also for the practicing clinician, and the medical teacher whose responsibility is to critically appraise the presented inferences before accepting them into practice or the curriculum. Development of new drugs, devices, and techniques is heavily dependent, nowadays, upon statistical analyses to prove their effectiveness. It is also frequently alleged that statistical analyses can be misused and misrepresented. This should be a further impetus for understanding inferential statistics.

Much of medical research begins with a research question that can be stated as a hypothesis. Testing of a hypothesis involves comparisons between sets of numbers. Comparisons are done for various reasons. A common reason for comparison is to see if there is a difference between data sets or groups. For instance, we may sample the price of 1 kg of potatoes and that of 1 kg of onions from different market locations and then compare the two sets of prices. We may be engaging in this exercise with the hypothesis that there is a price difference between the two commodities, with onions being the more expensive item. Suppose, we sample the price of onions per kilogram and that of 10 g gold from the same market locations. The intention in this latter exercise is unlikely to be finding out the price difference of the two items. However, one may have a hypothesis that, in Indian markets, the price of onions goes up whenever there is a hike in gold prices and hence the price sampling. The question in this case, therefore, is to check whether there is an association between commodity prices. Another reason for comparison may be to assess if there is an agreement between sets of numbers. Thus, in the same set of subjects, we may be measuring fasting blood glucose from capillary blood samples using a glucometer on one hand, and at the same time, from whole venous plasma samples sent to a reference

laboratory on the other. The intention here would be to compare the two sets of readings to judge if there is an agreement between them. Whatever the reason, comparisons are formally conducted as statistical tests of hypothesis and the results generated can be extrapolated to populations from which samples have been drawn.
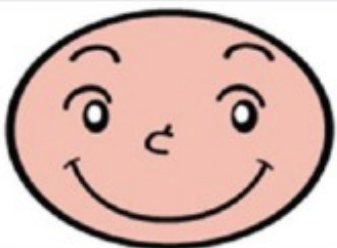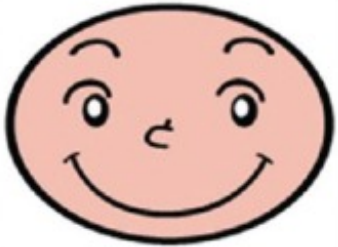
However, not all situations require formal hypothesis testing. Let us say, we give you two topical antifungal drugs and say that Drug A offers a cure rate of 20% in tinea corporis, while Drug B offers 80% cure rate in the same indication and ask which one will you use? Your answer, most likely, would be Drug B without hesitation. We do not require any formal hypothesis testing to draw an inference if the difference is this large. The simple descriptive statistics, as percentages, is enough to help us take a clinical decision. However, today if we are offered two antifungal drugs, the cure rates claimed are more likely to be around 95% and 97%. Which one to use then? This question is not answered so easily. Efficacy-wise, they appear to be comparable. We would, therefore, start looking at other factors such as adverse drug reaction profile, cost, and availability to arrive at a decision. We may even wonder if the observed 2% difference is a real occurrence or could it be just a chance difference? It is in such situations that hypothesis testing is required to help us arrive at a decision whether an observed change or difference is statistically 'significant' so that it may become the basis for altering or adapting clinical practice.

Go to:
The statistical convention in hypothesis testing is to begin with a null hypothesis that reflects the conservative position of no change in comparison to baseline or no difference between groups. Usually, the researcher has reason to believe that there is some effect or some difference - indeed this is usually the reason for the study in the first place! The null hypothesis is designated as H_0, while the clinician's working hypothesis may be one of a number of alternatives (e.g. $H_1 \equiv A > B$ or $H_2 \equiv B > A$). Hypothesis testing requires that researchers proceed to study samples and measure outcomes in the hope of finding evidence strong enough to be able to reject the null hypothesis. This is somewhat confusing at first but becomes easier to understand if we take recourse to the legal analogy. Under Indian criminal law, an accused is presumed to be not guilty unless proven otherwise. The task before the prosecution is therefore to collect and present sufficient evidence to enable to judge to reject the presumption of not guilty, which is the null hypothesis in this case.

While putting the null hypothesis to the test, two types of error may creep into the analysis: Type I error of incorrectly rejecting the null hypothesis, the probability of which is denoted by the quantity α, and Type II error of incorrectly accepting the null hypothesis, the probability of which is designated β. Again this is confusing and so let us return to the criminal court, where the verdict given may be at variance with reality. Let us say a person has not committed a murder but is accused of the same. The prosecuting side collects and presents evidence in such a manner that the judge pronounces the verdict as guilty. Thus, although the person has not committed the crime, the presumption of not guilty is falsely

rejected. This is Type I error. On the other hand, let us say that the person has committed the murder but the prosecution fails to provide enough evidence to establish guilt. Therefore, the judge pronounces the verdict as not guilty which is actually not the truth. This is Type II error. Note that although lay people and even the media may interpret not guilty as innocent, this is not the correct interpretation. The judge did not use the term "innocent" but gave the verdict as "not guilty." What the judge meant was that, in the technicalities of law, sufficient evidence was not presented to establish guilt beyond reasonable doubt. The statistician's stand in hypothesis testing is like that of the judge hearing a criminal case. He will start with the null hypothesis, and it is the clinician's job to collect enough evidence to enable the statistician to reject the null hypothesis of no change or no difference. diagrammatically depicts the concepts of Type I and Type II errors.

Diagrammatic representation of the concept of the null hypothesis and error types. Note that α and β denote the probabilities, respectively, of Type I and Type II errors. The happy faces represent error-free decisions

Statistician's also fondly speak of the power of a study. Mathematically, power is simply the complement of β, i.e., power = $(1-\beta)$. In probability terms, power denotes the probability of correctly rejecting H_0 when it is false and thereby detecting a real difference when it does exist. In practice, it is next to impossible to achieve 100% power in a study since there will always be the possibility of some quantum of Type II error. Type I and Type II errors bear a reciprocal relationship. For a given sample size, both cannot be minimized at the same time, and if we seek to minimize Type II error, the probability of Type I error will go up and vice

versa. Therefore, the strategy is to strike an acceptable balance between the two *a priori*. Conventionally, this is done by setting the acceptable value of α at no more than 0.05 (i.e., 5%) and that of β at no more than 0.2 (i.e., 20%). The latter is more usually expressed as a power of no less than 0.8 (i.e., 80%). The chosen values of α and β affect the size of the sample that needs to be studied - the smaller the values, the larger the size. For any research question, they are the two fundamental quantities that will influence sample size. We will deal with other factors that affect sample size in a latter module.

Go to:
The concept of the *P* value is almost universally used in hypothesis testing. Technically, *P* denotes the probability of obtaining a result equal to or "more extreme" than what is actually observed, assuming that the null hypothesis is true. The boundary for 'more extreme' is dependent on the way the hypothesis is tested. Before the test is performed, a threshold value is chosen called the significance level of the test (also denoted by α) and this is conventionally taken as 5% or 0.05. If the *P* value obtained from the hypothesis test is less than the chosen threshold significance level, it is taken that the observed result is inconsistent with the null hypothesis, and so the null hypothesis must be rejected. This ensures that the Type I error rate is at the most α. Typically the interpretation is:

- A small *P* value (<0.05) indicates strong evidence against the null hypothesis, so it is rejected. The alternative hypothesis may be accepted although it is not 100% certain that it is true. The result is said to be statistically "significant".

- An even smaller *P* value (<0.01) indicates even stronger evidence against the null hypothesis. The result may be considered statistically "highly significant".

- A large *P* value (>0.05) indicates weak evidence against the null hypothesis. Therefore, it cannot be rejected, and the alternate hypothesis cannot be accepted.

- A *P* value close to the cutoff (≈0.05 after rounding off) is considered to be marginal. It is better to err on the side of caution in such a case and not reject the null hypothesis.

Let us try to understand the *P* value concept by an example. Suppose a researcher observes that the difference in cure rate for pityriasis versicolor using single doses of two long-acting systemic antifungals on 50 subjects (25 in each group) is 11% with an associated *P* value of 0.07. This means that assuming the null hypothesis of no difference in a cure rate of the two antifungals to be true, the probability of observing a difference of 11% is 0.07, i.e. 7%. Since this is above the threshold of 5%, the null hypothesis cannot be rejected, and we have to go by the inference that the 11% difference may have occurred by chance. If another group repeats the study on 500 subjects (250 in each group) and observes the same 11% difference in cure rate, but this time with a *P* value of 0.03, the null hypothesis is to be rejected. The inference is that the difference in cure rate between the two drugs is statistically significant, and therefore the clinical choice should be with the more effective

drug. Note that, in this example, although the observed difference was the same, the *P* value became significant with the increase in sample size. This demonstrates one of the fallacies of taking the *P* value as something sacrosanct. The inference from the *P* value is heavily dependent on sample size. In fact, with a large enough sample, one can discover statistical significance in even marginal difference. Thus, if we compare two antihypertensive drugs on 10,000 subjects, we may find that an observed difference of just 2 mmHg in systolic or diastolic blood pressure also turns to be statistically significant. But is a difference of 2 mmHg clinically meaningful? Most likely it is not. If the observed difference is large, then even very small samples will yield a statistically significant *P* value. Therefore, *P* values must always be interpreted in the context of a given sample size. If the sample size is inadequate, studies will be underpowered, and even small *P* values will be clinically meaningless.

A *P* value is often given the adjective of being one-tailed (or one-sided) or two-tailed (or two-sided). Tails refer to the ends of a distribution curve, which typically has two tails. depicts the two tails of a normal distribution curve. A two-tailed *P* value implies one obtained through two-sided testing, meaning testing that has been done without any directional assumptions for the change or difference that was studied. If we are studying cure rates of two systemic antifungals for pityriasis versicolor, we should ideally do the testing as a two-sided situation since new Drug A may be better than existing Drug B and the reverse possibility also exists. It is seldom fair to begin with the presumption that Drug A can only be better than or similar to Drug B but will not be worse than it. However, consider a hypothetical drug that is claimed to increase height in adults. We may be justified in testing this drug as a one-tailed situation since we know that even if a drug cannot increase adult height, it will not decrease it. Decreasing height is a biological improbability unless the drug is causing bone degeneration in the axial skeleton. A one-tailed test is more powerful in detecting a difference, but it should not be applied unless one is certain that change or difference is possible only in one direction.
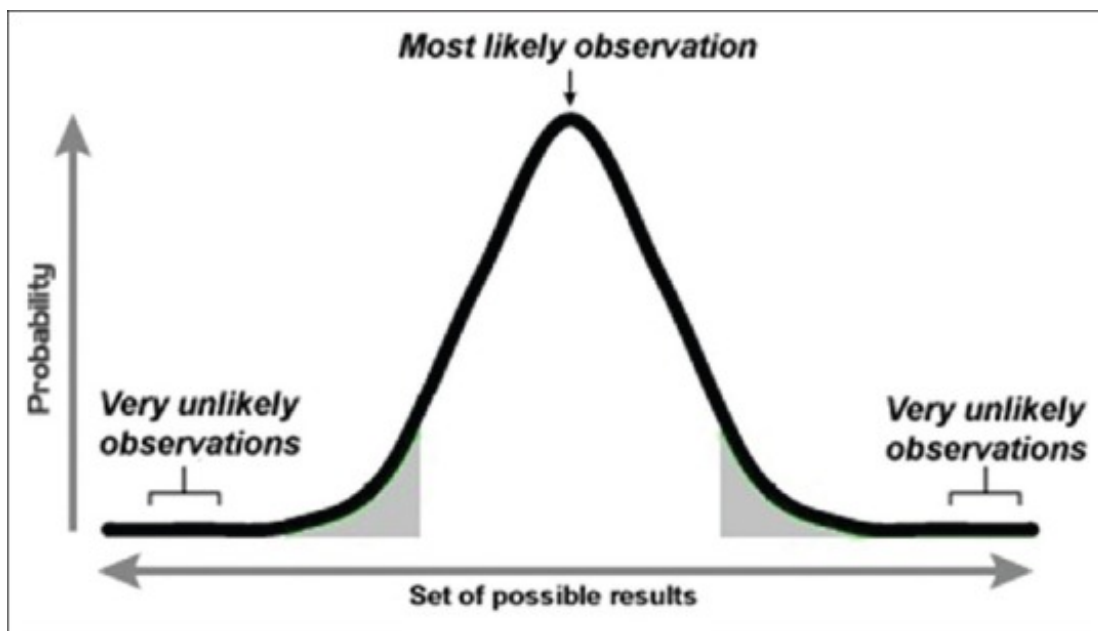
A normal distribution curve with its two tails. Note that an observed result is likely to return a statistically significant result in hypothesis testing if it falls in one of the two shaded areas, which together represent 5% of the total area. Thus, the shaded area is the area of rejection of the null hypothesis

Steps in hypothesis testing

Hypothesis testing, as it stands now, should proceed through the following five steps:

- Select a study design and sample size appropriate to the research question or hypothesis to be tested

- Decide upon the hypothesis test (i.e., test of statistical significance) that is to be applied for each outcome variable of interest

- Once the data have been collected, apply the test and determine the *P* value from the results observed

- Compare it with the critical value of P, say 0.05 or 0.01

- If the *P* value is less than the critical value, reject the null hypothesis (and rejoice) or otherwise accept the null hypothesis (and reflect on the reasons why no difference was detected).

With the increasing availability of computers and access to specialized statistical software (e.g. SPSS, Statistica, Systat, SAS, STATA, S-PLUS, R, MedCalc, Prism, etc.) the drudgery involved in statistical calculations is now a thing of the past. Medical researchers are therefore free to devote their energy to optimally designing the study, selecting the

appropriate tests to be applied based on sound statistical principles and taking care in conducting the study well. Once this is done, the computer will work on the data that is fed into it and take care of the rest. The argument that statistics is time-consuming can no longer be an excuse for not doing the appropriate analysis.

Which test to apply in a given situation

A large number of statistical tests are used in hypothesis testing. However, most research questions can be tackled through a basket of some 36 tests. Let us follow an algorithmic approach to understand the selection of the appropriate test. To do so, we convert a specific research question in our study to a generic question. It turns out that the vast majority of research questions we tackle can fit into one of five generic questions:
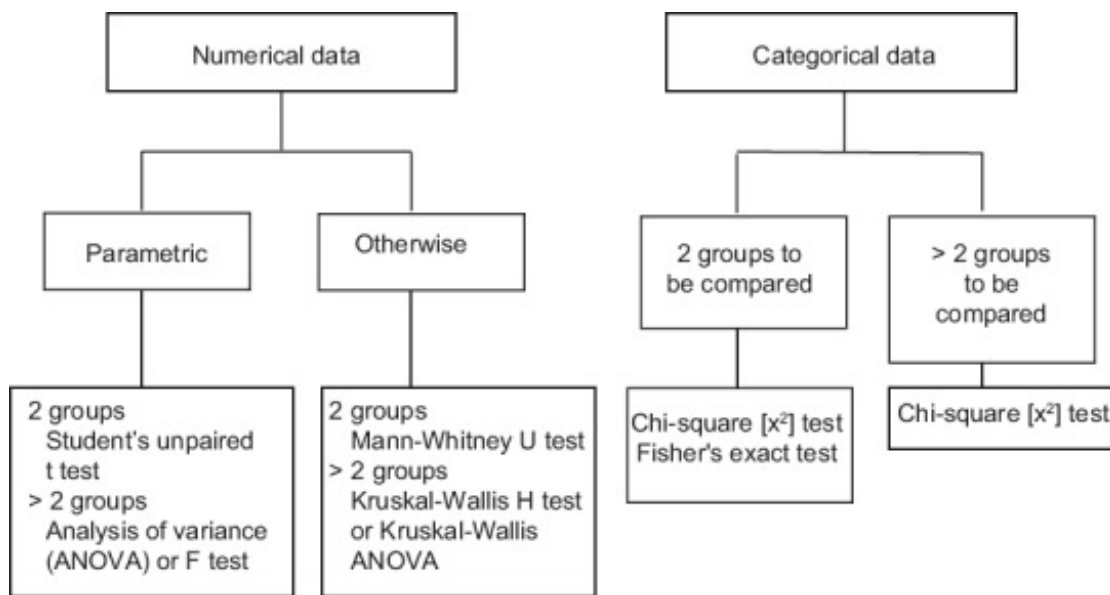
1. Is there a difference between groups or data sets that are unpaired?

2. Is there a difference between groups or data sets that are paired?

3. Is there an association between groups or data sets?

4. Is there an agreement between groups or data sets?

5. Is there a difference between time to event trends?

Let us pick up each question and follow the algorithm that leads to the tests. We will discuss the pros and cons of individual tests in subsequent modules. For the time being, let us concentrate on the schemes that are based on the context of individual questions.

Question 1. Is there a difference between groups or data sets that are unpaired (parallel or independent)?

This is perhaps the most common question encountered in clinical research. The tests required to answer this question are decided on the basis of nature of the data and the number of groups to be compared. The data sets or groups need to be independent of one another, i.e., there should be no possibility of the values in one data set influencing values in the other set or being related to them in some way. If related, then the data sets will have to be treated as paired. Thus, if we compare the blood glucose values of two independent sets of subjects, the data sets are unpaired. However, if we impose a condition like all the individuals comprising one data set are brothers or sisters of the individuals represented in the other data set, then there is possibility of corresponding values in the two data sets being related in some way (because of genetic or other familial reasons) and the data sets are no longer independent. provides a flowchart for test selection in the context of this question.

```
Numerical data                              Categorical data
    │                                            │
 ┌──┴──┐                                   ┌─────┴─────┐
 │     │                                   │           │
Parametric  Otherwise                  2 groups to   > 2 groups
                                       be compared    to be
                                                      compared
 │           │                             │           │
2 groups   2 groups                   Chi-square [x²]  Chi-square [x²] test
 Student's   Mann-Whitney U test       test
 unpaired  > 2 groups                  Fisher's exact
 t test      Kruskal-Wallis H test     test
> 2 groups   or Kruskal-Wallis
 Analysis of ANOVA
 variance
 (ANOVA) or F test
```

Tests to assess statistical significance of difference between data sets that are independent of one another

Note that numerical data have been subcategorized as "parametric" or "otherwise." Numerical data that follows the parameters of a normal distribution curve come in the first subcategory. In other words, parametric data are normally distributed numerical data. If the distribution is skewed, if there is no particular distribution or simply if the distribution is unknown, then the data must be considered as nonparametric. How do we know whether numeric data are normally distributed? We can look at the two measures of central tendency, mean and median. The properties of the normal distribution curve tell us that these should coincide. Hence, if mean and median are the same or are close to one another (compared to the total spread of data), then we are probably dealing with parametric data. However, a more foolproof way is to formally test the fit of the data to a normal distribution using one of a number of "goodness-of-fit" tests such as the Kolmogorov–Smirnov test, Shapiro–Wilk test, or D'Agostino and Pearson omnibus normality test. If such a test returns a $P < 0.05$, it implies that the null hypothesis of no difference between the data's distribution and a normal distribution will have to be rejected and the data taken to be nonparametric. The normal probability plot is a graphical method of deducing normality of continuous data.

Note also that whenever more than two groups are to be compared at a time we have a multiple group comparison situation. A multiple group comparison test (like one-way analysis of variance [ANOVA], Kruskal–Wallis ANOVA, or Chi-square test) will tell us whether there is a statistically significant difference overall. It will not point out exactly between which two groups or data sets does the significant difference lie. If we need to answer this question, we have to follow-up a multiple group comparison test by a so-called *post hoc* test. Thus, if ANOVA or its nonparametric counterpart shows a significant difference between multiple groups tested, we can follow them up with various *post hoc* tests like:

- Parametric data: Tukey's honestly significant difference test (Tukey–Kramer test), Newman–Keuls test, Bonferroni's test, Dunnett's test, Scheffe's test, etc.

- Nonparametric data: Dunn's test.

Question 2. Is there a difference between groups or data sets that are paired (cross-over type or matched)?

Data sets are considered to be paired if there is a possibility that values in one data set influence the values in another or are related to them in some way. There is often confusion over which data sets to treat as paired. The following provide a guide:

- Before-after or time series data: A variable is measured before an intervention, and the measurement is repeated at the end of the intervention. There may be periodic interim measurements as well. The investigator is interested in knowing if there is a significant change from baseline value with time

- A crossover study is done, and both arms receive both treatments, though at different times. Comparison needs to be done within a group in addition to between groups

- Subjects are recruited in pairs, deliberately matched for key potentially confounding variables such as age, sex, disease duration, and disease severity. One subject gets one treatment, while his paired counterpart gets the other

- Measurements are taken more than once, and comparison needs to be made between such repeated sets (e.g., duplicate or triplicate) of measurements

- Variables are measured in other types of pairs, for example, right-left, twins, parent-child, etc.

Many instances of pairing are obvious. A before-after (intervention) comparison would be paired. If we have time series data, such as Psoriasis Area and Severity Index scores estimated at baseline and every 3 months over 1 year, then all the five sets of data are paired to one another. Similarly, twin studies, sibling studies, parent-offspring studies, cross-over studies, and matched case–control studies, usually involve paired comparisons. However, some instances of pairing are not so obvious. Suppose we are evaluating two topical treatments for atopic dermatitis, which is a symmetrical dermatosis and decide to use one-half of the body (randomly selected) for the test treatment and the other half for some control treatment. We may assume that data accruing from the test half and the control half are independent of one another but is this assumption correct? Are we certain that if the test treatment works and the lesions on the test half regress, this is in no way going to influence the results in the control half? There would be absolutely no systemic

effect of the treatment that would influence the control half? If we are certain then by all means, we can go ahead and treat the two data sets as unpaired. Otherwise, it is preferable to treat them as paired.

provides the flowchart for test selection in the context of comparing data sets that show pairing. Note that the scheme remains the same as for the first question but the tests are different. Thus, Student's unpaired or independent samples *t*-test is now replaced by the Student's paired *t*-test, with its nonparametric counterpart now as the Wilcoxon's signed rank test. If we are comparing two independent proportions, such as the gender distribution in two arms of a parallel group clinical trial, we can use the Chi-square test or Fisher's exact test. However, if we are comparing paired proportions, such as the proportion of subjects with a headache before and after treatment of herpes zoster with aciclovir, then the test to employ is McNemar's test which is also a Chi-squared test.
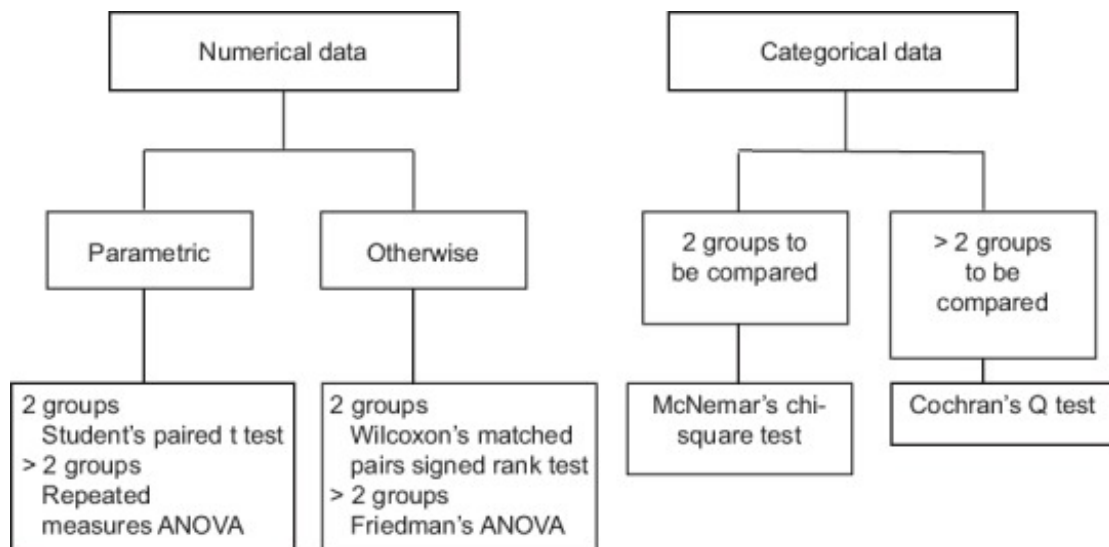

Tests to assess statistical significance of difference between data sets that are or could be paired

If multiple group comparisons are involved, such as repeated measures ANOVA or its nonparametric counterpart the Friedman's ANOVA, then once *again post hoc* tests are required if we are interested in deciphering exactly between which two data sets the significant difference lies. We have listed the *post hoc* tests under the first question. The same tests can be used adjusted for paired comparisons. However, note that *post hoc* tests are run to confirm where the differences occurred between groups, and they should only be run when an overall significant difference is noted. Running *post hoc* tests when the *a priori* multiple comparison test has not returned a significant *P* value is tantamount to 'data dredging' in the hope of discovering a significant difference by chance and then emphasizing it inappropriately.

Question 3. Is there an association between groups or data sets?

As seen from , the algorithm for deciding tests appropriate to this question is simpler. Correlation is a statistical procedure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates that the variables increase or decrease in parallel; a negative (or inverse) correlation indicates that as one variable increases in value, the other decreases.
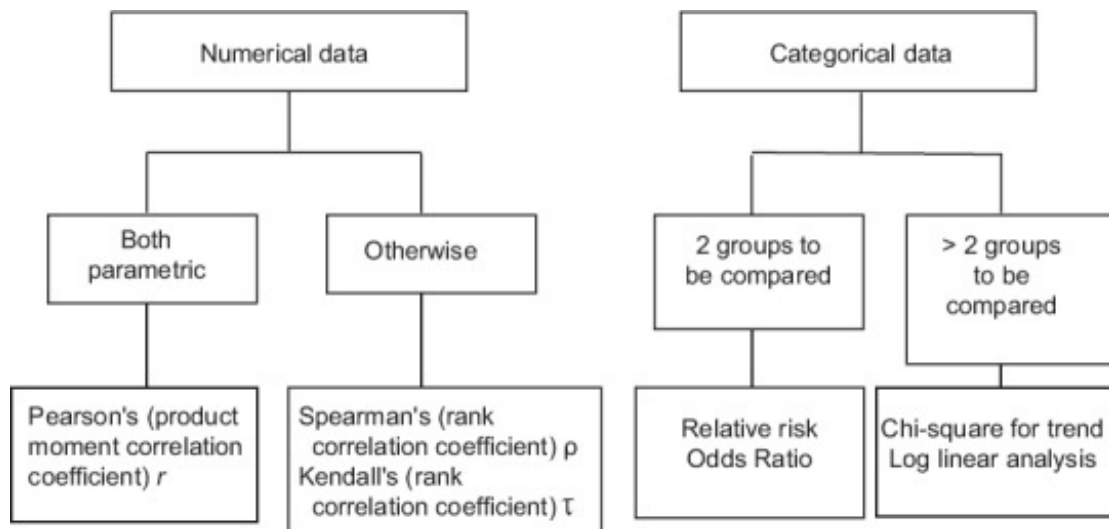
Tests for association between variables

With numerical data, we quantify the strength of an association by calculating a correlation coefficient. If both the variables are normally distributed, we calculate Pearson's product moment correlation coefficient r or simply Pearson's r. If one or both variables are nonparametric or we do not know what their distribution are, we calculate either Spearman's rank correlation coefficient Rho (ρ) or Kendall's rank correlation coefficient Tau (τ).

If numerical variables are found to be correlated to a strong degree, a regression procedure may be attempted to deduce a predictive quantitative relationship between them. In the simplest scenario, if two numerical variables are strongly correlated and linearly related to one another, a simple linear regression analysis (by least squares method) enables generation of a mathematical equation to allow prediction of one variable, given the value of the other.

Associating categorical data becomes simple if we can arrange the data in two rows and two columns as a 2 × 2 contingency table. Thus, if we want to explore the association between smoking and lung cancer, we can categorize subjects as smokers and nonsmokers and the outcome as lung cancer and no lung cancer. Arranging this data in a 2 × 2 table will allow ready calculation of a relative risk or an odds ratio, two measures of association that are hallmarks of epidemiological studies. However, if we categorize subjects as nonsmokers,

moderate smokers, and chain smokers, then we have to look for association by calculating a Chi-square for trend or other measures. We will take a detailed look at risk assessment in a future module.

Question 4. Is there an agreement between groups or data sets?

Agreement between groups or data sets can be inferred indirectly as lack of significant difference, but there are fallacies with this approach. Let's say in a university examination all candidates are being assessed independently by an internal examiner and an external examiner and being marked out of 50 in each case. In this instance, if we are interested in seeing to what extent the assessment of the two examiners agree, we may calculate the average marks (out of 50) given by the two examiners and see if there is a statistically significant difference between means by a paired comparison. If the difference is not significant, we may conclude that the two examiners have agreed in their assessments. However, we have compared group means. Even if group means tally, there may be wide variation in marks allotted to individual examinees by the examiners. What if the denominators were different - the internal examiner marking out of 30 and the external marking out of 70? The means will no longer tally, even if assessments of the performance of individual candidates are in agreement. We could of course convert to percentages and work with the mean of the percentages, but the initial fallacy remains. Measures of association were used earlier to denote agreement, but there are also problems there.

Therefore, statisticians agree that a better way of assessing agreement is to do so differently using measures that adjust individual agreements with the proportion of the disagreement in the overall result. The tests to do so are depicted in .
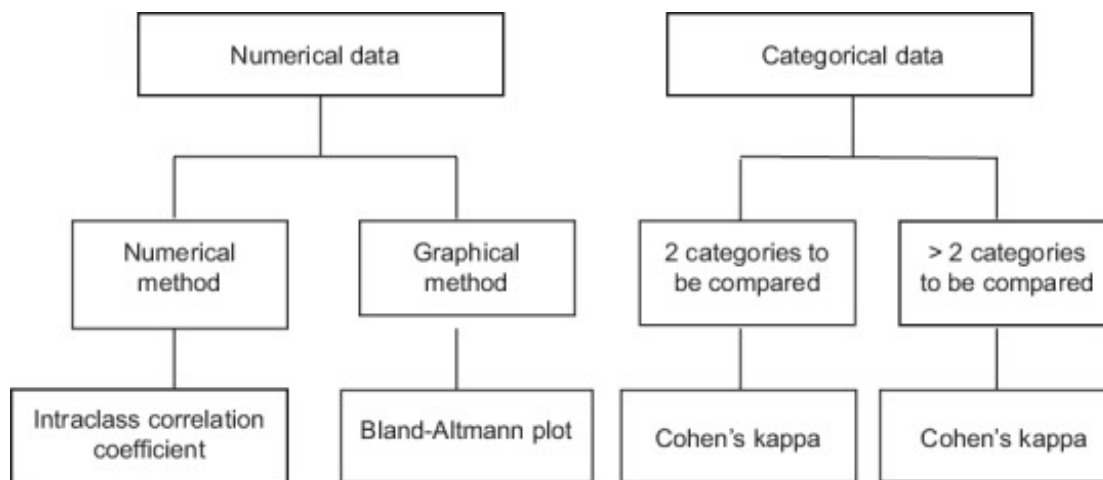

Tests for association between variables

An intraclass correlation coefficient, as we will see later in our module on correlation and regression, is interpreted in a manner similar to a correlation coefficient and is now employed widely in rater validation of questionnaires. Cohen's kappa statistic is used to

compare diagnostic tests that return their results as positive or negative or as some other categorical outcome.

Question 5. Is there a difference between time to event trends?

Time to an event that is going to happen in the (relatively distant) future is a special kind of numeric data. For example, let us consider two chemotherapy Regimens A and B for malignant melanoma of the skin. With Regimen A, we tell you that at the end of 5 years following treatment, 20% of the patients are expected to survive. With Regimen B, the expected 5 years survival rate is also 20%. Which regimen to choose? Obviously, on the face of it, both regimens are similar with respect to efficacy. Now, we divulge a little more information. Let's say with Regimen A there are no deaths in the first 3 years, then 40% of the patients die in the 4th year and another 40% in the 5th year, so 20% are surviving at the end of 5 years. With Regimen B, nobody dies in the 1st year but thereafter 20% of the patients die each year leaving 20% survivors at the end of 5 years. This time there is something to choose, probably Regimen A will be the natural choice. Thus, whenever we are considering the time to a future event as a variable, it is not only the absolute time, but also the time trend that matters. With such an example there are other complexities. For instance, if we start with hundred melanoma patients overall, then it may so happen that at the end of 5 years, twenty would be surviving, sixty would have died of cancer, ten may have died of unrelated causes, and the rest ten may be simply lost to follow-up. How do we reconcile these diverse outcomes when estimating survival? This issue is dealt with by censoring in survival studies.

The algorithm is simplest for this question, as seen in . We just have to decide on the number of groups to compare. In practice, the log-rank test is popularly used as it is simplest to understand and allows both two group and multiple group comparisons.
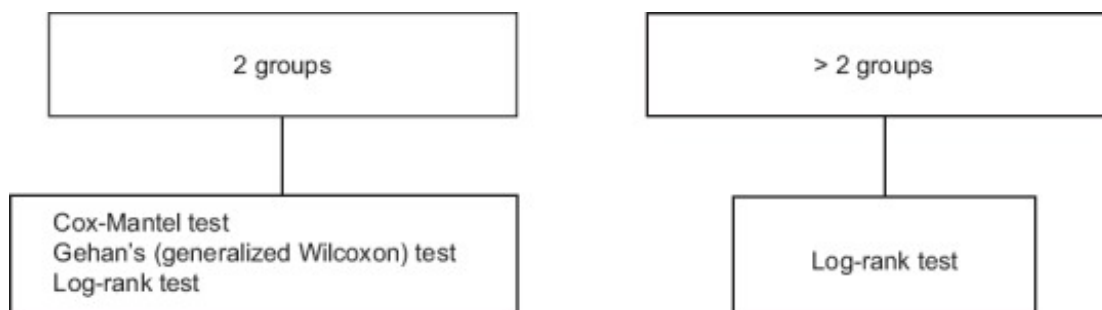
Tests for comparing time to event data sets

It is evident from the above schemes that when numerical data are involved, it is important to distinguish between parametric tests (applied to data that is normally distributed) and nonparametric tests (applied to data that is not normally distributed or whose distribution is unknown). It is sometimes possible to transform skewed data to a normal distribution (e.g.,

through a log transformation) and then analyze it through parametric tests. However, a better option is to use nonparametric tests for nonparametric data. Note that survival or time-to-event data are generally taken as nonparametric.

Some of the other assumptions of parametric tests are that samples have the same variance, i.e., drawn from the same population (Levene's test or Brown–Forsythe test can be applied to assess homogeneity of variances), observations within a group are independent and that the samples have been drawn randomly from the population. These assumptions are not always met but the commonly used tests are robust enough to tolerate some deviations from ideal.

Limitations of the *P* value approach to inference

The concept of the *P* value dates back to the 1770s when Pierre-Simon Laplace studied almost half a million births. He observed an excess of boys compared to girls and concluded by calculation of a *P* value that the excess was a real, but unexplained, effect. However, the concept was formally introduced by Karl Pearson, in the early 1900s through his Pearson's Chi-square test, where he presented the Chi-squared distribution and noted *p* as capital P. Its use was popularized by Ronald Aylmer Fisher, through his influential book Statistical Methods for Research Workers published in 1925. Fisher proposed the level $P < 0.05$ (i.e., a 1 in 20 chance of results occurring by chance) as the cut-off for statistical significance. Since then the *P* value concept has played a central role in inferential statistics, but its limitations must be understood.

Using the *P* value, researchers classify results as statistically "significant" or "nonsignificant," based on whether the *P* value was smaller than the prespecified cut-off. This practice is now frowned upon, and the use of exact *P* values (to say three decimal places) is now preferred. This is partly for practical reasons because the use of statistical software allows ready calculation of exact *P* values, unlike in the past when statistical tables had to be used. However, there is also a technical reason for this shift. The imputation of statistical significance based on a convention ($P < 0.05$) tends to lead to a misleading notion that a "statistically significant" result is the real thing. However, note that a *P* value of 0.05 means that 1 of 20 results would show a difference at least as big as that observed just by chance. Thus, a researcher who accepts a 'significant' result as real will be wrong 5% of the time (committing a Type I error). Similarly, dismissing an apparently "nonsignificant" finding as a null result may also be incorrect (committing a Type II error), particularly in a small study, in which the lack of statistical significance may simply be due to the small sample size rather than real lack of clinical effect. Both scenarios have serious implications in the context of accepting or rejecting a new treatment. It must be clearly understood that the *P* value is not the probability that the null hypothesis is true or the probability that the alternative hypothesis is false. The presentation of exact *P* values allows the reader to make an

informed judgment as to whether the observed effect is likely to be due to chance and this, taken in the context of other available evidence, will result in a more practical conclusion being reached.

Moreover, the *P* value does not give a clear indication as to the clinical importance of an observed effect. It is not like a score indicating that smaller the *P* value, more clinically important the result. A small *P* value simply indicates that the observed result is unlikely to be due to chance. However, just as a small study may fail to detect a genuine effect, a large study may yield a small *P* value (and a very large study a very small *P* value) based on a small difference that is unlikely to be of clinical importance. We discussed an example earlier. Decisions on whether to accept or reject a new treatment necessarily have to depend on the observed extent of change, which is not reflected in the *P* value.

P value vis-à-vis confidence interval

The *P* value provides a measure of the likelihood of a chance result, but additional information is revealed by expressing the result with appropriate confidence intervals. It is becoming the norm that an estimate of the size of any effect based on a sample (a point estimate), must be expressed with its 95% confidence interval (an interval estimate), for meaningful interpretation of results. A large study is likely to have a small (and therefore "statistically significant") *P* value, but a "real" estimate of the effect would be provided by the 95% confidence interval. If the intervals overlap between two treatments, then the difference between them is not so clear-cut even if $P < 0.05$. Conversely, even with a nonsignificant P, confidence intervals that are apart suggest a real difference between interventions. Increasingly, statistical packages are getting equipped with the routines to provide 95% confidence intervals for a whole range of statistics. Confidence intervals and *P* values are actually complementary to one another, and both have an important role to play in drawing inferences.

Go to:
The living world is full of uncertainty. Biostatistics is not a collection of clever mathematical formulae but is an applied science borne out of the need to make sense out of this uncertainty and in dealing with random variations that arise in any study situation. All medical personnel, whatever be their primary fields of activity, such as clinical practice, teaching, research, or administration, need to be aware of the fundamental principles of inferential statistics to correctly interpret study results and critically read medical literature to be able to make informed decisions furthering their own activity. This discussion has given an overview of hypothesis testing. To acquire working knowledge of the field, it is important to read through the statistical analysis part of research studies published in peer-reviewed journals. Mastery of the principles can, however, come only through actual number crunching using the appropriate statistical software.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

Go to:
1. Glaser AN, editor. High Yield Biostatistics. Baltimore: Lippincott Williams & Wilkins; 2001. Hypothesis testing; pp. 33–49.

2. Motulsky HJ, editor. Prism 4 Statistics Guide - Statistical Analysis for Laboratory and Clinical Researchers. San Diego: GraphPad Software Inc; 2005. *P* values and statistical hypothesis testing; pp. 16–9.

3. Motulsky HJ, editor. Prism 4 Statistics Guide - Statistical Analysis for Laboratory and Clinical Researchers. San Diego: GraphPad Software Inc; 2005. Interpreting *P* values and statistical significance; pp. 20–4.

4. Dawson B, Trapp RG, editors. Basic and Clinical Biostatistics. 4th ed. New York: McGraw Hill; 2004. Research questions about one group; pp. 93–133.

5. Dawson B, Trapp RG, editors. Basic & Clinical Biostatistics. 4th ed. New York: McGraw Hill; 2004. Research questions about two separate or independent groups; pp. 131–61.